3540–3545 *Nucleic Acids Research, 2003, Vol. 31, No. 13*
DOI: 10.1093/nar/gkg525

PromH: promoters identification using orthologous genomic sequences

V. V. Solovyev* and I. A. Shahmuradov[1]

Softberry Inc., 116 Radio Circle, Suite 400, Mount Kisco, NY 10549, USA and [1]Institute of Botany, Azerbaijan National Academy of Sciences, 370073 Baku, Azerbaijan

Received February 15, 2003; Revised and Accepted March 21, 2003

ABSTRACT

Accurate prediction of promoters is fundamental for understanding gene expression patterns, cell specificity and development. In the studies of conserved features of regulatory regions of orthologous genes, it was observed that major promoter functional components such as transcription start points, TATA-boxes and regulatory motifs, are significantly more conservative than the sequences around them (70–100% compared with 30–50%). To improve promoter identification accuracy, we employed these findings in a new program, PromH, created by extending the TSSW program feature set. PromH uses linear discriminant functions that take into account conservation features and nucleotide sequences of promoter regions in pairs of orthologous genes. The program was tested on two sets of pairs of orthologous, mostly human and rodent, sequences with known transcription start sites (TSS), annotated to have TATA (21 genes, 11 orthologous pairs) and TATA-less (38 genes, 19 pairs) promoters, respectively. The program correctly predicted TSS for all 21 genes of the first set with a median deviation of 2 bp from true site location. Only for two genes, was there significant (46 and 105 bp) discrepancy between predicted and annotated TSS positions. For 38 TATA-less promoters from the second set, TSS was predicted for 27 genes, in 14 cases within 10 bp distance from annotated TSS, and in 21 cases—within 100 bp distance. Despite more discrepancies between predicted and annotated TSS for genes from the second set, these results are consistent with observations of much higher occurrence of multiple TSS in TATA-less promoters. In any case, our results show that PromH identifies TSS positions significantly more accurately than any other published promoter prediction method. The PromH program is available at http://www.softberry.com/berry.phtml?topic=promh.

INTRODUCTION

The RNA polymerase II promoter is a key region that regulates differential transcription of protein coding genes. Gene-specific architecture of promoter sequences makes it extremely difficult to devise a general strategy for predicting promoters. Promoter 5′-flanking regions are especially poorly described and understood. They may contain dozens of short motifs (5–10 bases) that serve as recognition sites for proteins involved in transcription initiation and specific regulation of gene expression. Each promoter has unique selection and arrangement of such elements, which results in unique patterns of gene expression. There have been several reviews of promoter prediction approaches published recently (1–5).

The core promoter is a minimum promoter region that is capable of initiating basal transcription. It contains transcription start site (TSS) and typically spans from −60 to +40 relative to the TSS. About 30–50% of all known promoters contain TATA-box located ∼30 bp upstream of the TSS. TATA-box is apparently the most conserved functional signal in eukaryotic promoters and in some cases can direct accurate transcription initiation by POLII, even in the absence of other control elements. Many highly expressed genes contain strong TATA-box in their core promoter. However, in some large groups of genes, like housekeeping genes, oncogenes and growth factor genes, TATA box is often absent, and the corresponding promoters are referred to as TATA-less promoters. In these promoters, the exact position of the transcription start point may be controlled by nucleotide sequence of transcription initiation region (Inr) or recently found downstream promoter element (DPE), typically observed 30 bp downstream of the TSS (1,6,7).

The region 200–300 bp immediately upstream of the core promoter constitutes the proximal promoter. The proximal promoter usually contains multiple transcription factor binding sites, which are responsible for transcription regulation. The distal part of the promoter (usually known as enhancer/silencer elements) is located further upstream and may also include transcription factor binding sites (3–5).

The first comprehensive review of performance of many general-purpose promoter prediction programs was presented by Fickett and Hatzigeorgiou (8). Although their relatively

*To whom correspondence should be addressed. Tel: +1 914 242 3592; Fax: +1 914 242 3593; Email: victor@softberry.com

Present address:
I. A. Shahmuradov, Royal Holloway, University of London, Egham, Surrey TW20 0EX, UK

small test set (18 sequences) had several problems (9), the results demonstrated that tested programs can recognize just about 50% of promoters with a false positive rate of about 1 per 700–1000 bp (for more recent related reviews, see 3,9 and 10). Ohler *et al.* (9) used interpolated Markov chains in their approach and claimed to have slightly improved promoter prediction results, though they identified the same 50% of promoters from data set as Fickett and Hatzigeorgiou, while having one false positive prediction for every 849 bp. Later, to improve their own eukaryotic promoter recognition, Ohler *et al.* (11) applied an approach integrating some physical properties of DNA (DNA bendability, GC content) into their probabilistic promoter recognition system, McPromoter, and achieved a reduction of about 30% of false positives, compared with a model solely based on sequence likelihoods. The initial version of TSSW (12) had an accuracy of 42% with the false positive rate of 1 per 789 bp. Another promoter identification program, Promoter 2.0, was designed by Knudsen (13) applying combination of neural networks and genetic algorithms. Promoter 2.0 was tested on recognizing promoters in a complete adenovirus genome (35 937 bp). The program predicted all five known promoter sites on the plus strand and 30 false positive promoters. The average distance between actual and closest predicted promoter was about 115 bp. The TSSW program with the threshold to predict all five promoters produced 35 false positives, but its average distance between predicted and known TSS was just 4 bp (two promoters predicted exactly, one with 1 bp shift, one with 5 bp shift and the weakest promoter was predicted with 15 bp shift).

The current draft of human genome sequence provides a base for several annotations of genes, both known and predicted. These annotations, however, do not include promoters. Mapping known EST and mRNAs does not help: these sequences are usually 5′-incomplete. The first attempt to map promoter locations to the chromosome 22 sequence was based on the PromoterInspector program (14). The program can identify about 50% of known promoters as genomic regions up to 1 kb in length by discriminating them from the exon, intron and 3′-UTR sequences.

Recently, Bajic *et al.* (15) reported the Dragon Promoter Finder (DBF) program, which uses sensors for three functional regions: promoters, exons and introns, and an artificial neural network. Judging by authors' estimates, that approach has a higher accuracy than the three other compared promoter finding programs: NNPP2.1 (16), Promoter2.0 (13) and PromoterInspector (17). Another novel hybrid machine-learning method was reported by Down and Hubbard (18) that is able to predict >50% of human TSS with a false positive rate <30%.

Since gene prediction programs have a much greater accuracy than promoter prediction approaches, we suggested a different strategy of promoter identification. Instead of using classifiers to separate promoter regions from exon, intron and 3′-UTR sequences, we use upstream 5′-regions of annotated genes. Since 72% of such genes encode proteins similar to known proteins, these regions represent actual 5′-sequences in at least 72% of cases. To define promoter location in query sequences, we use a modification of the TSSW program, which can localize promoter and TSS within several nucleotides from actual location. As shown in the elegant work of Wasserman *et al.* (19) for human and mouse orthologous genes (genes in two species that have arisen from the same locus of their common ancestor) upregulated in skeletal muscle, 98% of experimentally defined transcriptional factor binding sites are confined to 19% of human DNA sequence that is the most conserved. We used several types of conservative blocks to enhance sensitivity and specificity of the TSSW algorithm, providing pairs of aligned orthologous genomic sequences as input data. Recently, the draft sequence of mouse genome (20) and a gene expression map of human chromosome 21 orthologues in the mouse (21) have been reported. Therefore, this strategy can be applied for promoter prediction of most human and mouse genes.

In this paper, we present the results of our analysis of conservative elements in pairs of orthologous genes of human and related species, and then we show how implementation of these results can be used to improve promoter identification. Finally, we give comprehensive instructions to potential users of PromH.

RESULTS

Analysis of conservative elements in regulatory regions of orthologous genes

There are many research articles that study promoter elements that are conservative in particular pairs of orthologous genes of different species (see 19 and references therein). In this work, we tried to find features that are universal for all PolII promoters and can be used on any pair of orthologous genes. For analysis we took sequences of two groups of genes annotated to have, mostly, TATA and TATA-less promoters, respectively: 21 genes of 'TATA' group (11 human; four otolemur; three mouse; three rat or three human–mouse; three human–rat and five human–otolemur pairs) and 38 genes of 'TATA-less' group (19 human genes and their 17 mouse and two rat orthologues, totaling 19 pairs). Exhaustive information on these genes, including their annotated promoter type and TSS is given in the Supplementary Material (Tables S1 and S3, respectively). According to available annotations, at least 15 of 21 genes from the first set have known TATA-promoters. For the other six genes, promoter type ($TATA^+$ or TATA-less) is not defined. As to the second set, excluding only one gene (human *c-erb2/new*) annotated to have TATA-promoter, all genes have annotated TATA-less promoters. Depending on the availability of the DNA sequence, the upstream regions of these genes (2000..CDS start position) have been analyzed. The total lengths of analyzed sequences in the first and second sets are 38 363 and 63 578 bp, respectively.

The full-length sequences of gene pairs have been aligned by the SCAN2 program (http://softberry.com/berry.phtml?topic=scanh&prg=SCAN2), which can align megabases of genomic regions in seconds and can work with several sets of parameters accounting for strong and weak similarity, the latter being very useful for accurate alignment of promoter sequences in orthologous genes. The alignments revealed that general base identity between upstream regions of related genes from the first set is relatively weak: for four pairs, ~30%; for five pairs, 40–50% and only for one pair (human and rat MYL3 genes), 61%. In the second set, six pairs have

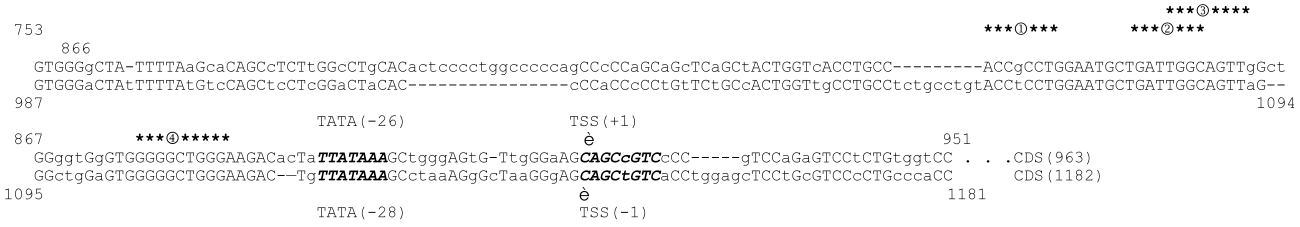

Figure 1. The location of predicted TSS and TATA boxes as well as some of the conserved regulatory motifs in aligned sequences of *h-PGAM-M* and *r-PGAM2* orthologous gene pairs of human (top line) and rat. TSS and TATA boxes are given in bold italic. Conservative regulatory motifs from the Transfac database (23) are denoted by encircled numbers: 1, HS$TPI_04; 2, PA$PY_21; 3, MOUSE$M2EB; 4, SP1$CONS. For additional information, see Table S1 in the Supplementary Material.

Table 1. Comparing TSS prediction results for two sets of genes by TSSW and PromH

	Set 1: 21 genes (TATA promoters)		Set 2: 38 genes (TATA-less promoters)	
	True	False-negative	True	False-negative
TSS predicted by TSSW	15	6	22	16
TSS predicted by PromH	21	0	27	11

Predicted TSSs in the area from −150 to +10 for TATA promoters and from −300 to +100 for TATA-less promoters (in relation to the annotated TSS), were considered as true or reasonable predictions.

similarity of ≤37%; for seven pairs it is within 40–60% range; and for six pairs, ≥60% or more. At the same time, alignment reveals many blocks with a very high level of conservation. Interestingly, these interspecies conservative blocks are not distributed equally along upstream regions. This finding supports a large volume of literature data indicating a key role of only some domains (blocks) of upstream gene regions in regulation of transcription. Therefore, we expected that such conservative 'islands' would help in more accurate prediction of promoters.

For TATA promoters (set 1) we found four classes of such blocks making meaningful contribution to predicting 'true' promoters.

1. *Conservative region of TATA-box (14 bp including TATA-box 4 positions).* Seventeen out of 20 'true' TATA-promoters have an interspecies conservation level in the region >70% (six of them are identical). For the other three promoters, the conservation levels were 62% (compared to 28% for the whole 5′-region), 47% (31%) and very low 25% (30%).
2. *Conservative region of TSS (9 bp around TSS).* Thirteen out of 21 genes have ≥77% level of similarity (five are identical), six genes have 66%, one gene has 41% and only one gene has 25% conservation level.
3. *Above average conservation level of regulatory motifs to the left from TSS region.* Sixteen out of 21 genes have homology >70%, the remaining five have 45–56% [in this analysis, collection of regulatory elements from TRANSFAC (22; http://transfac.gbf.de/TRANSFAC/) was used].
4. *Conservative region to the right from TSS.* Thirteen out of 21 genes have >70% homology in this region, seven have >50%, and one has 45%.

A similar conservation level for regions 2–4 is also observed in TATA-less promoters.

Features in discriminant functions of TSSW and PromH programs

The TSSW program (8) classifies each position on a given sequence as TSS or non-TSS based on two linear discriminant functions (for TATA and TATA-less promoters) with eight characteristics calculated in the (−200, +50) region around the current position. If the TATA-box weight matrix gives a score higher than some preliminary defined threshold in the region (−40, −25) from current position, then that position is classified based on LDF for TATA promoters, otherwise it will be classified by LDF for TATA-less promoters. For any pair of predicted TSS, located within 300 bp from each other, only the one with the highest LDF score is retained, except for one case: if a lower scoring position is predicted by the LDF for TATA-less promoters near a higher scoring position predicted by LDF for TATA-promoters, then the first position is also retained as a potential enhancer region.

To take advantage of the knowledge of conserved elements in 5′-regions of homologues genes, we added the following new features to the list of features already used in discriminant function for distinguishing promoters and non-promoters in TSSW: conservativeness levels of regions (i) around TSS and (ii) to the right of TSS (40 bp), (iii) an average conservation level of regulatory motifs located to the left of TSS and, for TATA promoters, (iv) conservation level around TATA-box. These features were implemented in a novel promoter (TSS) prediction program PromH (H stands for homology).

In the case of the prediction of more than one promoter (TSS) with high LDF score and if CDS start is known, the TSS

```
     Program  PromH  (Softberry Inc.)
     Search for TATA+/TATA- promoters in 2 aligned DNA sequences

 NOTE: PHa  - Homology Level of Aligned Sequences in LOCAL Search Area
       PHs  - Homology Level of Aligned Sequences around TSS
       PHss - Homology Level of Aligned Sequences to Right from TSS
       PHt  - Homology Level of Aligned Sequences around TATA-box
       PHr  - Mean Homology Level of Regulatory Elements in LOCAL Search Area
==================================================
>h-PGAM2 [1:962]/-920:61/ AC J05073
 Length of sequence-       981
 Initial / Final Thresholds for TATA+ promoters -   0.10 /  2.50
 Initial / Final Thresholds for TATA-/enhancers -   0.70 /  3.70
     2 promoter/enhancer(s) have been predicted
 Enhancer Pos:    899 (LDF:   5.79)
     PHa -  68%   PHs - 100%   PHss -  22%   PHr -  76%
 Promoter Pos:    921 (LDF-  3.61)   TATA box at:    895 (LDF-    18.51)
     PHa -  66%   PHs -  77%   PHss -  23%   PHt -  70%   PHr -  71%

 Transcription factor binding sites:
for promoter at position -     921
   752 (+) MAIZE$ADH1   CGTGG
   631 (+) Y$ADH2_01    TCTCC
   854 (+) HS$ALBU_02   TTGGCA
   853 (+) MOUSE$A21C   ATTGG
   824 (+) MOUSE$MCK_   cccaaCACCTGCtgcctgagcc
   ....................
--------------------------------------------------
>r-PGAM2 [-1181..+800: 1:2160] AC Z17319/ TATA: 1120..1128/CDS: 1182..2245)
 Length of sequence-      1300
 Initial / Final Thresholds for TATA+ promoters -   0.10 /  2.50
 Initial / Final Thresholds for TATA-/enhancers -   0.70 /  3.70
     2 promoter/enhancer(s) have been predicted
 Enhancer Pos:   1123 (LDF:   3.97)
     PHa -  68%   PHs - 100%   PHss -  22%   PHr -  80%
 Promoter Pos:   1148 (LDF-  2.83)   TATA box at:   1119 (LDF-    17.83)
     PHa -  65%   PHs -  88%   PHss -  23%   PHt -  70%   PHr -  82%

 Transcription factor binding sites:
for promoter at position -    1148
   902 (+) Y$ADH2_01    TCTCC
   935 (+) HS$ALBU_02   TTGGCA
  1081 (+) MOUSE$A21C   ATTGG
   942 (+) RAT$EAI_08   ccctgccCAGCTGgc

.....................................................
```

Figure 2. Results of the promoter search in human *h-PGAM-M* and rat *r-PGAM2* orthologous genes by PromH (example of finding only TATA promoters). Only the first few regulatory motifs are presented.

```
     Program  PromH  (Softberry Inc.)
     Search for TATA+/TATA- promoters in 2 aligned DNA sequences

 NOTE: PHa  - Homology Level of Aligned Sequences in LOCAL Search Area
       PHs  - Homology Level of Aligned Sequences around TSS
       PHss - Homology Level of Aligned Sequences to Right from TSS
       PHt  - Homology Level of Aligned Sequences around TATA-box
       PHr  - Mean Homology Level of Regulatory Elements in LOCAL Search Area
==================================================
>pr22h..pr(22)/human,AC004595/h-SP4 gene/PROM:2960 nt(2572-2647..TSSs???)
 Length of sequence-      2960
 Initial / Final Thresholds for TATA+ promoters -   0.10 /  2.50
 Initial / Final Thresholds for TATA-/enhancers -   0.70 /  3.70
     4 promoter/enhancer(s) have been predicted
 Promoter Pos:   2639 (LDF:  21.46)
     PHa -  84%   PHs - 100%   PHss -  17%   PHr -  86%
 Promoter Pos:   1950 (LDF:  11.37)
     PHa -  80%   PHs -  77%   PHss -  -2%   PHr -  94%
 Promoter Pos:   2336 (LDF:   4.94)
     PHa -  79%   PHs - 100%   PHss -   9%   PHr -  63%
 Promoter Pos:    391 (LDF-  3.19)   TATA box at:    358 (LDF-    19.99)
     PHa -  61%   PHs -  91%   PHss -  37%   PHt -  75%   PHr -  55%

 Transcription factor binding sites:
for promoter at position -    2639
  2478 (-) HS$A4_01     GGGCGCgGG
  2583 (+) CHICK$ACRA   CCGCCC
  2594 (+) CHICK$ACRA   CCGCCC
  2613 (+) CHICK$ACRA   CCGCCC
.....................................................

--------------------------------------------------
>pr22m..pr(22)/mouse,AB019147/m-SP4 gene/PROM:3020 nt(2656-2731..16 TSSs)
 Length of sequence-      3020
 Initial / Final Thresholds for TATA+ promoters -   0.10 /  2.50
 Initial / Final Thresholds for TATA-/enhancers -   0.70 /  3.70
     4 promoter/enhancer(s) have been predicted
 Promoter Pos:   2685 (LDF:  26.76)
     PHa -  82%   PHs -  77%   PHss -  16%   PHr -  86%
 Promoter Pos:   2052 (LDF:  11.37)
     PHa -  82%   PHs -  88%   PHss -  -1%   PHr -  96%
 Promoter Pos:    135 (LDF-  2.93)   TATA box at:    104 (LDF-    18.75)
     PHa -  45%   PHs -  83%   PHss -  38%   PHt -  39%   PHr -  40%
 Promoter Pos:   1672 (LDF-  2.56)   TATA box at:   1640 (LDF-    18.15)
     PHa -  72%   PHs -  88%   PHss - -19%   PHt -  90%   PHr -  68%

 Transcription factor binding sites:
for promoter at position -    2685
  2564 (-) HS$A4_01     GGGCGCgGG
  2667 (+) CHICK$ACRA   CCGCCC
  2678 (+) CHICK$ACRA   CCGCCC
  2553 (-) CHICK$ACRA   CCGCCC
.....................................................
```

Figure 3. Results of the promoter search in human *h-SP4* and mouse *m-SP4* orthologous genes by PromH (example of finding both TATA and TATA-less promoters). Only the first several regulatory motifs are presented.

closest to the CDS start was assumed as a predicted promoter. But, of course, it might be a choice of user.

Promoter prediction by TSSW and PromH programs

The results of promoter prediction for the TATA-promoter set of genes by the PromH program are summarized in Table S1 of the Supplementary Material. PromH has found 20 of 21 interspecies conservative TATA-promoters (including their TATA-boxes and TSS). Positions of the predicted TSS coincide with annotated pre-mRNA start positions or differ from them by 1–5 bp, and the average discrepancy between predicted and annotated TSS is just 2 bp (just for one predicted TSS, the discrepancy was 105 bp). Regulatory motifs and main components (TATA-box and TSS) of predicted TATA-promoters are very conservative in orthologous genes and these predictions correspond closely to the available promoter annotations (Fig. 1; for other examples of TATA-promoter predictions and their conservative blocks see Figs S1 and S2 in the Supplementary Material).

Of all the TATA-promoters, only for the mouse *GLUT4* gene, was a high-score TATA-less promoter located close to annotated TSS predicted.

At the same time, in a few genes, such as *m-GLUT4*, *h-GLUT4* and *h-NPPA*, some discrepancy between predicted and annotated TSS localization is observed. Such discrepancies may have several explanations. The GenBank annotation for the *m-GLUT4* gene includes a putative weak TATA-box, which was never experimentally supported. Our detailed analysis of this region did not find any motif resembling the consensus of the TATA box. More detailed analysis of human and mouse orthologous *GLUT4* gene pairs shows that the upstream regions of each gene contain two potential promoters with a very high score (see Fig. S2 in the Supplementary Material): *h-GLUT4*: (i) TSS: −105, LDF = 4.01; (ii) TSS: −459, LDF = 5.96; *m-GLUT4*: (i) TSS: −46, LDF = 10.23; (ii) TSS: −405, LDF = 3.21.

The conservation level around all four putative (two for every gene) TSS is also very high (∼70%). Therefore, we cannot exclude that in this case both promoters are functional. Additional study of occurrences of potential alternative promoters revealed a similar situation for the *NPPA* genes of human and rat (data not shown). Taking into consideration results of this analysis, it may be concluded that for all 21 genes, 'true' promoters have been predicted by PromH.

For comparison, 15 out of 21 TSS mentioned above were also predicted by TSSW (Table 1; see also Table S2 in the Supplementary Material). At the same time, TSSW did not predict any TSS for genes (*ol-HBE*, *r-MLCIV*, *r-PGAM2* and *ol-HBGG*). Finally, for two genes (*h-MYF4* and *m-MYOG*) distances between annotated TSS and those predicted by TSSW were very large, 1066 and 862, respectively.

On the second set of mostly TATA-less promoters (this set was not used in learning significant characteristics for PromH), we observed that TSS were predicted in 27 out of 38 genes (see Table S3 in the Supplementary Material), including 14 genes with 0–10 bp distance between predicted and annotated TSS or 21 genes with the corresponding distance < 100 bp. For four genes, the distance was between 101 and 150 bp and for two genes it was >150 bp. Despite a significant discrepancy between predicted and annotated TSS for genes from the second set, these findings might reflect the fact that the TSS of TATA-less promoters are more flexible than those of TATA-promoters (they often have multiple TSS and, usually, are extremely poorly predicted by the available programs). Interestingly, for four genes, instead of the annotated TATA-less promoters we found TATA-promoters with a high conservation level of TATA-box: *h-NCX* (TATA-box conservation level, 100%), *h-RPE65* (80%), *m-HIC-1* (45%) and *h-neu4* (77%). Compared to TSSW, PromH is significantly more accurate in predicting TATA-less promoters as well (Table 1; see also Table S4 in the Supplementary Material).

WEB availability of PromH and its input and output information

The PromH program is available at http://www.softberry.com/berry.phtml?topic=promh. To run this program online, two sequences of upstream regions of orthologous genes in FASTA format are required. Initially, downloaded query sequences are aligned by SCAN2. Then a promoter search is performed on aligned sequences, and promoter prediction results are displayed on the screen. Examples of PromH output for TATA and TATA-less promoters are presented in Figs 2 and 3, respectively.

DISCUSSION

Our results indicate that comparing orthologous genomic sequences substantially improves the quality of promoter identification. In contrast to previous works that mainly focused on groups of specific genes, we found several universal characteristics independent of gene type, which can be used by a general promoter prediction program. These characteristics derived from alignments of orthologous genes using SCAN2 program with parameters to search for weak but significant similarities. SCAN2 was specifically designed for comparing genomic sequences, so it aligns a pair of 10 000-bp 5′-regions in a second.

Integration of all available information about chromosome locus speeds up the research process of the analysis and cloning of new genes. In many cases, starting with just a fragment of EST with an interesting pattern of expression from micro-array experiments, the scientist can map it to the human genome sequence and computationally produce the exon structure of a corresponding gene. Predicted genes simplify primer selection for full-length cDNA cloning, because most exons are predicted correctly. One of the best resources of such integrated information is the Human Genome Browser developed at UCSC by Jim Kent: http://genome/cse.ucsc.edu/goldenPath/octTracks.html. Another interactive Java-based Genome Browser was designed by Softberry Inc. and contain most of the public and some proprietary annotation data: http://www.softberry.com/berry.phtml?topic=ge-hg13. These sites provide information about the known and predicted gene location and mapping of known mRNA and EST sequences, which in many cases provide independent support of location of transcribed gene sequence. This information alone allows us to locate 5′-regions of most human genes within 0.2–10 kb fragment. Using these regions, known mouse syntenic sequences and TSSW or PromH programs, we can identify most transcription starts correctly. Recently we have implemented the PromH approach using alternative transcription factor database (TFD) (23).

We should note that it is the first attempt to use general promoter information from orthologous genes to precisely define the start of transcription. More studies in this direction should be made to confirm our findings on the larger set of gene pairs. Also, analysis of more difficult to predict promoters, such as TATA-less promoters, alternative promoters, promoters of genes with several non-coding exons, which can have TSS dozen thousand bp upstream of the protein-coding region, should be carried out.

Our work demonstrates that using orthologous sequences, we can not only predict promoter regions (0.2–1 kb) as in the earlier work (9), but relatively accurately, within 1–5 nucleotides, locate start of transcription. Moreover, PromH provides not only TSS positions, but also locations of known regulatory elements around it. Taking into account that currently we have just several hundred experimentally supported promoter sequences in human genome, PromH program may be of considerable value for molecular biologists in deciphering regulation of genes encoded in sequenced genomes or interpreting results of expression profiling.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

REFERENCES

1. Smale,S.T. (1997) Transcription initiation from TATA-less promoters within eukaryotic protein-coding genes. *Biochim. Biophys. Acta*, **1351**, 73–88.
2. Werner,T. (1999) Models for prediction and recognition of eukaryotic promoters. *Mamm. Genome*, **10**, 168–175.
3. Pedersen,A.G., Baldi,P., Chauvin,Y. and Brunak,S. (1999) The biology of eukaryotic promoter prediction—a review. *Comp. Chem.*, **23**, 191–207.
4. Lemon,B. and Tjian,R. (2000) Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev.*, **14**, 2551–2569.
5. Smale,S.T. (2001) Core promoters: active contributors to combinatorial gene regulation. *Genes Dev.*, **15**, 2503–2508.
6. Burke,T.W. and Kadonaga,J.T. (1997) The downstream promoter element, DPE, is conserved from *Drosophila* to humans and is recognized by TAFII60 of *Drosophila*. *Genes Dev.*, **11**, 3020–3031.
7. Perier,C.R., Praz,V., Junier,T., Bonnard,C. and Bucher,P. (2000) The eukaryotic promoter database (EPD). *Nucleic Acid Res.*, **28**, 302–303.
8. Fickett,J. and Hatzigeorgiou,A. (1997) Eukaryotic promoter recognition. *Genome Res.*, **7**, 861–878.
9. Ohler,U., Harbeck,S., Niemann,H., Noth,E. and Reese,M. (1999) Interpolated Markov chains for eukaryotic promoter recognition. *Bioinformatics*, **15**, 362–369.
10. Ohler,U. and Niemann,H. (2001) Identification and analysis of eukaryotic promoters: recent computational approaches. *Trends Genet.*, **17**, 56–60.

11. Ohler,U., Niemann,H., Lia,G.-C. and Rubin,G.M. (2001) Joint modeling of DNA sequence and physical properties to improve eukatyotic promoter recognition. *Bioinformatics*, **17**, S199–S206.
12. Salamov,A.A. and Solovyev,V.V. (1997) The Gene-Finder computer tools for analysis of human and model organisms genome sequences. In Rawling,C., Clark,D., Altman,R., Hunter,L., Lengauer,T. and Wodak,S. (eds), *Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology*, AAAI Press, Halkidiki, Greece, pp. 294–302.
13. Knudsen,S. (1999) Promoter 2.0: for the recognition of PolII promoter sequences. *Bioinformatics*, **15**, 356–361.
14. Scherf,M., Klingenhoff,A., Frech,K., Quandt,K., Schneider,R., Grote,K., Frisch,M., Gailus-Durner,V., Seidel,A., Brack-Werner,R. and Werner,T. (2001) First Pass Annotation of promoters of human chromosome 22. *Genome Res.*, **11**, 333–340.
15. Bajic,V.B., Seah,S.H., Chong,A., Zhang,G., Koh,J.L.Y. and Brusic,V. (2002) Dragon promoter Finder: recognition of vertebrate RNA polymerase II promoters. *Bioinformatics*, **18**, 198–199.
16. Reese,M., Harris,N.L. and Eeckman,F.H. (1996) Large scale sequencing specific neural networks for promoter and splice site recognition. In Hunter, L. and Klein, T.E. (eds), *Biocomputing Proceedings of the 1996 Pacific Symposium, 2–7 January, 1996*. World Scientific, Singapore, pp. 737–738.
17. Scherf,M., Klingenhoff,A. and Werner,T. (2000) Highly specific localization of promoter regions in large genomic sequences by Promoter Inspector: a novel context analysis approach. *J. Mol. Biol.*, **297**, 599–606.
18. Down,T.A. and Hubbard,T.J. (2002) Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res.*, **12**, 458–461.
19. Wasserman,W.W., Palumbo,M., Thompson,W., Fickett,J.W. and Lawrence,C.E. (2000) Human-mouse genome comparisons to locate regulatory sites. *Nature Genet.*, **26**, 225–228.
20. Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
21. The HSA21 Expression Map Initiative (2002) A gene expression map of human chromosome 21 orthologues in the mouse. *Nature*, **420**, 586–590.
22. Wingender,E., Chen,X., Hehl,R., Karas,H., Liebich,I., Matys,V., Meinhardt,T., Reuter,M., Reuter,I. and Schacherer,F. (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.*, **28**, 316–319.
23. Ghosh,D. (1993) Status of the transcription factors database (TFD). *Nucleic Acids Res.*, **21**, 3117–8318.